Depth-Aware Generative Adversarial Network for Talking Head Video Generation

Year: 2022 | Citations: 210 | Authors: Fa-Ting Hong, Longhao Zhang, Li Shen, Dan Xu

Abstract

Talking head video generation aims to produce a synthetic human face video that contains the identity and pose information respectively from a given source image and a driving video. Existing works for this task heavily rely on 2D representations (e.g. appearance and motion) learned from the input images. However, dense 3D facial geometry (e.g. pixel-wise depth) is extremely important for this task as it is particularly beneficial for us to essentially generate accurate 3D face structures and distinguish noisy information from the possibly cluttered background. Nevertheless, dense 3D geometry annotations are prohibitively costly for videos and are typically not available for this video generation task. In this paper, we introduce a self-supervised face-depth learning method to automatically recover dense 3D facial geometry (i.e. depth) from the face videos without the requirement of any expensive 3D annotation data. Based on the learned dense depth maps, we further propose to leverage them to estimate sparse facial keypoints that capture the critical movement of the human head. In a more dense way, the depth is also utilized to learn 3D-aware cross-modal (i.e. appearance and depth) attention to guide the generation of motion fields for warping source image representations. All these contributions compose a novel depth-aware generative adversarial network (DaGAN) for talking head generation. Extensive experiments conducted demonstrate that our proposed method can generate highly realistic faces, and achieve significant results on the unseen human faces. 11https://github.com/harlanhong/CVPR2022-DaGAN